第三方惩罚对合作的溢出效应：基于社会规范的解释*

陈思静 [1]　邢懿琳 [1]　翁异静 [1]　黎常 [2]

（[1]浙江科技学院经济与管理学院　杭州　310023）（[2]浙江工商大学工商管理学院　杭州　310018）

摘要 第三方惩罚对合作的维系可能来自经济功能或规范提示功能。先前研究没有区分这两种功能，因而未能回答：当惩罚不足以影响违规收益时，是否还能促进合作？实验一（$N = 252$）发现即使第三方惩罚无法降低违规收益，依然能抑制自利行为。实验二（$N = 179$）发现受过惩罚的违规者在其后的独裁者博弈表现出了更高的合作水平。2（是否旁观惩罚）×2（旁观前后）设计的实验三（$N = 179$）显示，旁观惩罚后被试的合作水平显著高于旁观前，也高于未旁观惩罚的被试。后两个实验中，社会规范在惩罚与合作之间均起中介作用。这进一步证实惩罚对合作的促进在很大程度上是通过规范激活来实现的，并存在两种溢出效应：惩罚抑制了曾经的违规者（纵向溢出效应）和旁观者（横向溢出效应）在新博弈情境下的自私行为。这两种溢出效应的发现补充了文献中占主导地位的经济学解释，并为理解人类社会长时间、大规模的合作提供了新视角。

关键词 第三方惩罚，社会规范，合作，聚焦理论，溢出效应

分类号 B849: C91

1 前言

在社会科学中，合作指的是个体付出成本使他人受益的行为（Nowak, 2006; Rand, 2016），非亲缘个体间的广泛合作对人类社会的顺利运行至关重要（Fehr & Schurtenberger, 2018），为此我们发展出了合作的社会规范（de Kwaadsteniet et al., 2007），即被群体成员所普遍接受但不同于法律条规等明文规章的有关合作的行为准则（Cialdini & Trost, 1998）。尽管合作规范普遍存在于各个文化中，但对合作规范的遵守并非自然而然之事（de Kwaadsteniet et al., 2019），而第三方惩罚（third-party punishment）——由利益无关者针对违规者所实施的惩罚——总体上被认为是减少违规行为并维系合作规范的重要力量之一（Balliet et al., 2011; Fehr & Gächter, 2002; Halevy & Halali, 2015）。在此基础上，学者探讨了规范在惩罚影响合作过程中的作用，如 Bicchieri 等（2018）发现，惩罚需要与一定的社会规范相结合才能发挥积极作用；类似地，Fehr 和 Williams（2018）也注意到，只有当群体成员间存在相应的规范共识时，

收稿日期：2020-06-08

* 国家自然科学基金项目（71701185），浙江省软科学项目（2020C35020），浙江省自然科学基金项目（LQ18G010002）资助。

通信作者：黎常，Email: lichang@zjgsu.edu.cn

第三方惩罚才能起到促进合作的正面作用，当规范共识缺席时，惩罚反而加速了社群的崩溃；此外，Lois 和 Wessa（2019）还探讨了社会规范对第三方惩罚的调节作用。但另一个更为基本的问题是第三方惩罚为什么能减少（促进）违规（合作）行为，而我们注意到，在回答这个问题上，基于规范视角的研究是相对缺席的。目前，对上述问题一种广为接受的解释主要基于经济学视角，即第三方惩罚改变了违规者的收益结构：存在第三方惩罚的情况下，个体的违规成本将大幅上升以至超过违规行为所带来的收益（韦倩，姜树广, 2013; Bicchieri et al., 2018; Carpenter & Matthews, 2004; Nelissen & Mulder, 2013; Rand et al., 2010），在这种情况下，理性个体的占优策略是选择合作而非违规。

然而，上述基于经济学视角的解释可能存在若干问题。第一，大量研究表明人们在决策过程中并不总是遵循经济人原则（Alkan, 2020; Camerer & Fehr, 2006; Henrich et al., 2001），因此，除非我们先入为主地预设违规者恰好总是纯粹理性的经济人，否则单纯从经济角度很难充分解释第三方惩罚对违规的抑制作用，而这个预设是否合理尚有探讨空间。第二，先前有研究者发现惩罚者的动机显著影响了惩罚的作用（谢东杰，苏彦捷, 2019; Raihani & Bshary, 2015），如 Rand 等（2009）指出，惩罚是否被认为合理可以极大地影响受罚者的反应；而 Fehr 和 Rockenbach（2003）也注意到，当惩罚被认为是出于自利（比如惩罚是为了获取更多的个人利益），尽管惩罚能显著降低违规收益（减少的金额等于初始金额的 40%），但受罚的违规者并没有表现出更高的合作水平，结果恰恰相反，其合作水平明显下降了。如果惩罚促进合作主要是由于其降低了违规收益，那么上述发现便难以得到合理的解释。第三，如果惩罚对违规的抑制作用主要在于提高了违规成本，那么有理由认为，除非在任何情况下违规都会受到惩罚，否则曾经受罚的经历不足以使个体在新情境下自动表现得更好。然而，正如 Shreedhar 等（2018）指出，如果一个群体必须对任何违规都实施惩罚，这个群体将付出极为高昂的代价，这部分代价甚至超过了惩罚所来的积极作用。换言之，无处不在的惩罚不仅无法维持大规模社群中的合作行为，反而会导致这类群体在竞争中失去优势。

基于上述原因，我们认为纯粹的经济学观点不足以充分解释第三方惩罚对合作规范的维系作用。陈思静等（2015）基于社会规范聚焦理论（focus theory of normative conduct）（Cialdini et al., 1991）提出第三方惩罚本身即是一种社会规范的激活过程，这为我们更好地理解第三方惩罚提供了另一种理论起点。社会规范聚焦理论认为，人们做出违规行为可能只是没有意识到存在某种规范，因此，只要通过某种方式让规范成为人们的意识焦点，便可以显著降低人们的违规行为。事实上，有研究者基于上述角度发现第三方惩罚确实能起到激活社会规范的作用（陈思静等, 2015），而 Chen 等（2020）也注意到，第三方惩罚能够显著地影响人们

的规范感知。然而，在先前研究中第三方惩罚通常改变了违规者的收益结构，这意味着先前研究者未能严格区分第三方惩罚的两种功能：通过降低违规收益来提升合作（惩罚的经济效应）以及通过激活社会规范来提升合作（惩罚的规范效应）。本文拟在这方面为现有文献提供有益补充，具体而言，本文将于实验 1 中在控制违规者收益的情况下检验第三方惩罚的规范激活功能。如果实验结果显示，尽管惩罚并未降低违规者的收益，但受罚的违规者依然表现出了较高的合作行为，那么我们就可以在一定程度上认为，惩罚的规范效应是一种独立于经济效应的功能，并且为社会规范聚焦理论提供了新的实证证据：激活人们的规范就可改变其行为。

其次，人类社会的合作表现出长时间和大规模的特点（Bingham, 1999），而如果惩罚的作用仅仅体现为受罚者本人在某个特定场景下的合作规范被激活从而提高了合作水平，那么我们又陷入了类似用经济学观点去解释合作的理论困境：假如必须通过惩罚对每个个体在每个场景下进行规范提示，那么社会的运行成本会变得极高，从而使第三方惩罚失去存在的意义（Shreedhar et al., 2018）。因此，我们推测第三方惩罚的规范提示作用不仅体现为抑制了违规者当下的自私行为，而且这一规范激活的效应还可以延续至新的场景（纵向溢出效应或时间维度上的溢出效应，实验 2）以及目睹而非亲身经历惩罚的旁观者（横向溢出效应或空间维度上的溢出效应，实验 3），即使在这两种情况下并不存在潜在的惩罚者。如果上述推测成立，那我们就可以在一定程度上解释为什么真实生活中并非时时刻刻发生了第三方惩罚，但人类社会的合作依然得以有条不紊开展的原因。

最后，社会规范作为被群体成员广泛接受并区别于法律规章的行为准则（Cialdini & Trost, 1998; Forquesato, 2016），在社会科学文献中通常被区分为描述性规范（descriptive norm）和命令性规范（injunctive norm）（Cialdini et al., 1991）：前者指的是人们在某一方面的普遍行为模式，如合作的描述性规范可理解为人们所表现出来的合作行为的普遍程度；而后者指的是人们对某一行为普遍所持赞成或批评的态度，如合作的命令性规范可理解为人们对他人合作行为的赞成程度。社会规范可显著影响人们的行为，如简化个体的行为决策并使个体在面对复杂、不确定甚至是危险的情境时得到行为上的指引（McDonald & Crandall, 2015）。但需要说明的是，研究者从不同角度指出了两种规范在影响行为中的区别，如 Deutsch 和 Gerard（1955）指出人们对描述性规范的认知加工速度要高于对命令性规范的加工，因此，描述性规范通常更容易对行为产生影响；而 Petty 和 Cacioppo（1986）从个人卷入度（personal involvement）比较了两种规范对行为的影响，并指出当个人卷入度较高时，命令性规范的作用更大。就本文而言，一个值得探讨的问题是当惩罚通过激活社会规范来影响合作时，惩罚

是激活了其中一种规范还是两种规范都有所激活？如果两种规范都被激活了，那么它们是否具有不同的作用机制？我们将在实验 2 和 3 中详细探讨这些问题。此外，由于社会规范聚焦理论的重点考察对象是描述性规范，如果我们的实验结果表明，在惩罚通过激活规范而影响合作的过程中，命令性规范也被激活并产生了显著影响，那么本文的结果也可在一定程度上被视为对这一理论的有益补充。

基于对上述文献的回顾，我们提出以下研究问题作为本文的主要探索目标：

研究问题 1：当第三方惩罚无法降低违规者收益时，是否依然能有效减少（促进）违规（合作）行为？（实验 1）

研究问题 2：第三方惩罚通过规范激活而提升合作的作用是否能溢出到新的情境？（实验 2）

研究问题 3：第三方惩罚通过规范激活而提升合作的作用是否能溢出旁观者身上？（实验 3）

研究问题 4：描述性和命令性规范在惩罚通过规范激活影响合作的过程中是否具有相似的作用机制？（实验 2 和 3）

概括而言，本文拟从社会规范的视角来解释第三方惩罚对合作的影响机制：我们认为规范激活是第三方惩罚的一种独立功能，即便无法降低违规收益，第三方惩罚依然可以抑制（促进）个体的违规（合作）行为（实验 1），同时，这一效应还溢出到了缺乏惩罚机制的新场景中（实验 2）和目睹惩罚行为的旁观者上（实验 3）。此外，我们还检验了两种规范在上述过程中的作用机制（实验 2 和 3），并讨论了这些发现的理论和现实意义。

2 实验 1：惩罚的规范效应

2.1 被试

取中等效应量 $f = 0.25$，显著性水平 $\alpha = 0.05$，通过软件 G* Power3.1 进行的功效分析（power analysis）显示，3 组间单因素方差分析至少需要 252 名被试才能达到 95%（$1 - \beta$）的统计检验力。考虑到本实验采用了“4+1”的实验设计，每 5 名被试中有 4 名被试的数据是进行统计分析的有效数据（详情见 2.2 部分）。我们共招募了 315 名来自浙江工商大学不同专业的本科生。所有被试在实验开始前详细阅读了实验说明并签署了知情同意书。实验正式开始前我们通过若干练习题使被试熟悉了实验规则（例题见附录）。用于统计分析的 252

名有效被试平均年龄为 21.42±2.25 岁，其中女性占比为 58.33%，被试的专业分布如下：理工科占 34.92%、社会科学占 28.57%、人文学科占 25.40%、艺术及其他占 11.11%。

2.2 设计与程序

实验 1 为 3（对照组、高收益组和低收益组）组间因子设计。实验 1 的范式为公共物品博弈，通过 z-Tree 上机实验的方式完成（Fischbacher, 2007）。实验期间，被试位于单独隔间内，相互间无法交流。实验 1 中每 5 人组成一个小组进行博弈，其中 4 人为参与者，参与公共物品博弈，剩余 1 人为执行者[1]，执行者不参与博弈，其在对照组中扮演收税人的角色，而在其他两种实验条件下则扮演惩罚人。为了排除直接互惠（direct reciprocity）（Trivers, 1971）、间接互惠（indirect reciprocity）（Nowak & Sigmund, 1998）和高成本信号（costly signaling）（Gintis et al., 2001）等机制的潜在影响，在每一轮博弈中，4 名参与者被随机编号为 A、B、C、D，而执行者的编号始终是 E，小组成员都由计算机随机安排，但参与者和执行者的角色不能互换。每一轮博弈结束时告知被试该组每个成员在该轮博弈中的贡献和收益（在有惩罚条件下，反馈还包括惩罚情况），但是在新一轮的博弈中，被试并不知晓同组成员在过去博弈中的表现。另外，为了避免尾轮效应（end effect），被试事前并不知晓博弈轮数。

实验开始后，被试被随机平均分入 3 种实验条件：对照组（n＝84）、高收益组（n＝84）和低收益组（n＝84）。在对照组中，每个被试（包括参与者 A/B/C/D 和执行者 E）在实验开始前拥有 25 代币（相当于 5 人民币）的初始金额。在每一轮博弈中，参与者自由决定是否将初始金额中的 10 代币投入公共账户中，而投入公共账户的代币翻倍后平均分给小组所有参与者。对于参与者而言，占优策略是选择保留初始金额（违规）并尽可能让他人将代币投入公共账户，而非将 10 个代币投入公共账户（合作）。但是，如果每人都这样做的话，最终每个人的收益反而降低了。此外，参与者被告知，选择保留 10 代币的话则需要向成员 E 缴纳 1 代币的收入税，该金额不进入公共账户，也不返还给任何成员。完成上述步骤后随即进入下一轮，总共进行 10 轮。10 轮博弈后实验结束，实验者对被试进行反馈和支付报酬。实验报酬为 10 元出场费加上随机抽取一轮被试手中的代币数（5 代币可换 1 人民币，下同）。

高收益组和低收益组的实验流程和对照组基本类似，主要区别在于这两种实验条件下，在计算机反馈参与者的选择后，成员 E 可惩罚违规者[2]：1）高收益组中，如果执行者选择惩

---

[1] 为了避免某些词语可能带有的感情色彩对被试产生影响，在实验中“参与者”、“执行者”或“惩罚”等词均被“角色 A/B/C/D”、“角色 E”和“扣减”所取代。

[2] 在实验 1 中，被试只能惩罚违规者，即选择保留 10 代币的参与者，这一设置是为了避免反社会惩罚（antisocial punishment）——针对合作者的惩罚（Herrmann et al., 2008）——对实验结果的干扰。

罚某个（些）参与者，那么他/她每惩罚一次需支付 5 代币，而受罚者只需支付 1 代币作为违规成本，因此违规收益较高；2）在低收益组中，执行者支付 5 代币而受罚者支付 10 代币作为违规成本，因此违规收益较低。在每一轮中，执行者可同时惩罚多名违规者，但对每一位违规者只能惩罚一次。在被试完成惩罚决定后，计算机公布上述决定及每个被试在本轮中的收益。表 1 总结了三种实验条件下参与者在某轮博弈中选择合作/违规的收益。

表 1 不同实验条件下合作与违规的收益

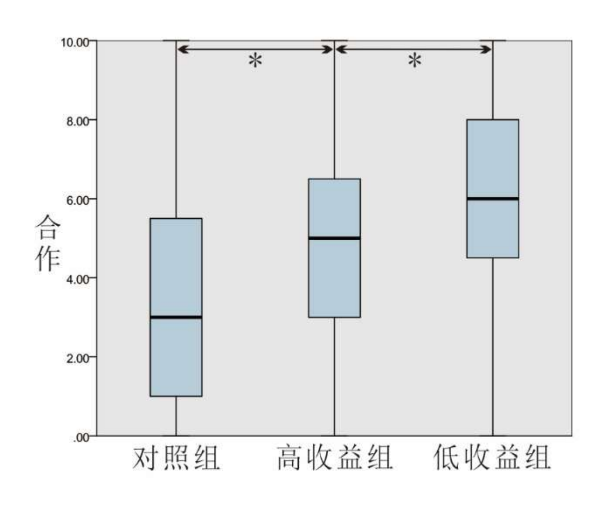
实验条件	U_C	U_D
对照组	$25+\frac{2（10x_C+10）}{4}-10$	$25+\frac{20x_C}{4}-1$
高收益组	$25+\frac{2（10x_C+10）}{4}-10$	$25+\frac{20x_C}{4}-x_D$
低收益组	$25+\frac{2（10x_C+10）}{4}-10$	$25+\frac{20x_C}{4}-10x_D$

注：U_C表示选择合作的收益；U_D表示选择违规的收益；x_C表示其他选择合作的人数（$x_C\in\{0,1,2,3\}$）；x_D表示被惩罚的次数（$x_D\in\{0,1\}$）。

实验 1 重点在于比较高收益组和对照组被试的合作水平（平均每轮投入公共账户的钱数）。从表 1 可以看出，这两组的合作收益是一样的，差别在于违规收益，并且高收益组的违规收益大于等于对照组，因为公式中 $X_D\leq1$，根据纯粹理性人的观点，相比于对照组，高收益组被试更有动力去选择违规，我们应该能观察到高收益组的合作水平低于对照组。其次，通过比较低收益组被试和高收益组被试的合作水平，我们可以在一定程度上考察经济因素对抑制违规行为的作用，因为这两组被试的唯一差别在于低收益组被试的违规成本远高于高收益组（前者是后者的 10 倍）。

2.3 结果与讨论

不同性别（$t=0.83, p=0.408$）和专业（$F=1.54, p=0.204$）下合作水平的差异不显著，年龄与合作水平（$r=-0.03, p=0.597$）的相关系数不显著。运用单因素方差分析比较三组被试的合作水平，结果显示，三组被试的合作水平存在显著差异（F = 15.24，p < 0.001，d = 0.65，95%C.I. = [0.38，0.92]）。多重比较（Tukey 法）的结果表明：高收益组被试（M = 4.75，SD = 2.57，n = 84）的合作水平显著高于对照组（M = 3.55，SD = 2.80，n = 84）（p = 0.012，95%C.I. = [0.22，2.19]），而低收益组被试（M = 5.86，SD = 2.76，n = 84）的合作水平显著高于对照组（p < 0.001，95%C.I. = [1.32，3.30]）和高收益组（p = 0.023，95%C.I. = [0.12，2.09]）。图 1 直观地显示了三组被试合作水平的差异。



注：$^{*}p < 0.05$

图 1 三组被试的合作水平

上述结果一方面验证了降低违规收益对提高合作水平的重要性，也就是说，在惩罚抑制违规行为的过程中，基于成本-收益的经济考虑确实发挥了显著的作用。这一点体现在了低收益组被试的合作水平显著高于高收益组，这意味着通过第三方惩罚改变违规者的收益结构确实可以激励人们减少违规行为从而提升了合作水平（Balliet et al., 2011; Gächter et al., 2008）。但从另一方面来说，和我们预期相似，经济因素无法完全解释实验 1 的结果。对比表 1 中高收益组和对照组违规收益（U_D）可以看出，高收益组被试选择违规的预期收益总是大于等于对照组，因此，根据经济人逻辑，高收益组中将有更多（少）的被试选择违规（合作）。然而事实上，高收益组被试合作水平却显著高于对照组，这说明相较于对照组而言，即使第三方惩罚并没有本质上降低高收益组违规者的收益，但依然可以有效地抑制违规(促进合作)。这意味着惩罚降低违规行为的心理机制不仅仅是其改变了违规者的收益结构，一定还存在其他重要因素。换句话说，人们在决策过程中并非总是遵循经济人假设这一原则同样可应用在违规者身上，尽管这个结果可能违反了我们的直觉。

通过回答研究问题 1，即第三方惩罚对合作的促进并不完全取决于其降低违规收益的作用，实验 1 也在一定程度上支持了聚焦理论的观点：很多时候人们做出违规行为并不是单纯为了追求利益，而只是没有意识到某种规范的存在（Cialdini et al., 1991）。就实验 1 而言，对比高收益组和对照组的实验条件，唯一的差别在于违规成本：高收益组为 X_D，而对照组为 1，$X_D \leq 1$，但 X_D 对违规的抑制作用却更高，因此，有理由认为抑制违规作用的差别主要来自两种成本的质的差异而非量的差异：表现为惩罚的违规成本提示了人们对违规行为持道德批评的态度，从而激活了人们有关合作的社会规范（陈思静等, 2015），而对照组中表现为收入税的违规成本却相对中性，缺乏这一功能。另外，尽管先前也有研究（e.g., 陈思静等,

2015; Chen et al., 2020）提出惩罚具有规范提示的功能，但由于在大部分情况下，惩罚总是会影响受罚者的经济利益，因而无法在严格意义上回答下列问题：当第三方惩罚不足以改变违规者的收益结构时，是否还能有效地促进合作？实验 1 首次通过随机对照实验控制了经济收益对实验结果的影响，从而为第三方惩罚的规范提示作用提供了明确的实证证据，这意味着惩罚的规范效应并非需要经济效应为前提，这对现有研究的结论是一个有力的补充。

3 实验 2：惩罚促进合作的纵向溢出效应

实验 1 为第三方惩罚纯粹的规范提示功能提供了证据，实验 2 进一步检验惩罚提升合作的功能是否能溢出到不存在惩罚机制的新情境下，并比较描述性和命令性规范的作用机制，从而回答本文所提出的研究问题 2 和 4。

3.1 被试

来自不同专业的 300 名学生参加了实验 2，并在实验开始前详细阅读了实验说明并签署了知情同意书。实验 2 需要首先筛选出违规者。根据实验 1 差异比较的结果 $d = 0.65$，取显著性水平 α = 0.05，用 G*Power3.1 计算出实验 2 至少需要由 104 名违规者组成的样本才能达到 95%（1 – β）的统计检验力，而通过实验 2 阶段一的操作，我们总共得到了 179 名违规者。这 179 名被试平均年龄为 21.30 ± 1.97 岁，其中女性占比为 54.19%，被试的专业分布如下：理工科占 35.75%、社会科学占 31.84%、人文学科占 24.02%、艺术及其他占 8.38%。

3.2 设计与程序

3.2.1 第一阶段：有第三方的独裁者博弈

实验 2 为 2（对照组 vs. 惩罚组）组间因子设计。实验 2 的范式为带有第三方的独裁者博弈。在阶段一中，被试被告知他/她将与其他 2 名被试组成一个小组来完成 5 轮独裁者博弈。在 5 轮博弈中，被试均扮演分配者，而扮演接受者和第三方的 2 名被试实际是虚拟被试，即由实验者事先设定的计算机程序[3]。此外，被试还被告知每一轮博弈开始前，分配者、接受者和第三方分别拥有 10、0 和 2 代币的初始金额，分配者可将初始金额在其和接受者之间自由分配，而接受者无权反对，但第三方可对其认为不公平的方案进行惩罚，惩罚规则为第三方付出 2 代币扣减分配者 6 代币。另外，被试还通过指导语了解到在每一轮博弈中，小组成员都是由计算机随机选择的，并且每轮博弈均无结果反馈。在实际操作中，基于先前文献的结论（Csukly et al., 2011; Fehr & Fischbacher, 2003），判断被试的分配方案是否违规的标

[3] 在实验指导语中分配者、接受者和第三方分别用角色 A、角色 B 和角色 C 代替，下同。

准如下：当被试分配给接受者的金额小于初始金额的 30%时，分配方案即被判定为违规，反之即为合作。完成上述 5 轮博弈后，共有 179 名被试在 5 轮博弈中至少有过一次违规行为，这些被试在阶段二中被随机分入两组：90 被试被告知其在过去 5 轮博弈中受到了来自第三方的惩罚（惩罚组），而剩余的 89 名被试则没有任何反馈（对照组）[4]。

3.2.2 第二阶段：独裁者博弈和公共物品博弈

分组后，对照组和惩罚组被试完成以下任务：1）与其他 1 名被试共同完成 1 轮无第三方的独裁者博弈，在博弈中他们将继续扮演分配者，但分配方法与前一阶段有所不同：每个被试拥有 20 代币的初始金额，他们可自由选择初始金额的一部分（0~10 之间的任一整数）分配给接受者，并且被试被明确告知不管他/她的分配方案如何，都不会遭受惩罚；2）与其他 3 名被试共同完成 1 轮无第三方的公共物品博弈，在博弈中他们可自由地将 20 代币初始金额的一部分（0~20 之间的任一整数）投入公共账户，投入公共账户的金额将翻倍后在 4 名成员中平均分配，并且被试被明确告知不管他/她的选择如何，都不会遭受惩罚。为了避免顺序对结果的潜在影响，一半被试先阅读有关独裁者博弈的指示语，另一半被试顺序相反。接着，被试分别估计在独裁者博弈中：1）将 0、1、2…10 代币分配给接受者的被试的百分比；2）赞成将 0、1、2…10 代币分配给接受者的被试的百分比；3）从 0~10 选择一个整数代表自己愿意分配给接受者的金额；以及在公共物品博弈中：4）从 0~20 选择一个整数代表自己愿意投入到公共账户的金额。完成上述步骤后，实验者宣布实验结束，并对被试进行反馈和支付报酬。实验报酬为 10 元出场费加上随机抽取一轮被试手中的代币数。

我们用两种方式来测量被试在博弈中描述性和命令性规范的激活水平：第一种采用 Chen 等（2020）的方法，用 1）和 2）这两项各自的加权平均值分别作为描述性规范和命令性规范激活水平的操作定义；在第二种方法中，我们采用 Bicchieri 和 Xiao（2009）、Voisin 等（2016）以及 Sood 等（2020）的范式，即使用被试对某个行为或赞成某个行为普遍程度的估计来代表被试的描述性或命令性规范的激活水平，具体而言，即有多少比例的分配者将（赞成将）20 代币中的 7、8、9 和 10 代币分配给接受者，以此作为两种规范激活水平的操作定义[5]。在统计分析中我们主要采用第一种操作定义来检验研究问题，并采用第二种操作

[4] 在阶段一中没有违规行为的 121 名被试不再参与下一阶段实验，但为了保证实验的顺利进行，这些被试被告知接下去他们将完成一轮旨在测试“外语思维对规范感知的影响”的实验，具体任务为阅读一份由英语写作的关于非洲某部落礼物交换规范的短文并回答相应问题。

[5] 先前有相当文献表明在不同文化语境中人们对于什么样的分配方案算是违规/合作有高度稳定的看法，即分配给对方的金额约小于 30%是一种违规行为（Csukly et al., 2011, Fehr & Fischbacher, 2003），且有学者认

定义作为稳健性检验，考察在两种操作定义下结果是否有质的差别，从而增强研究结论的说服力。

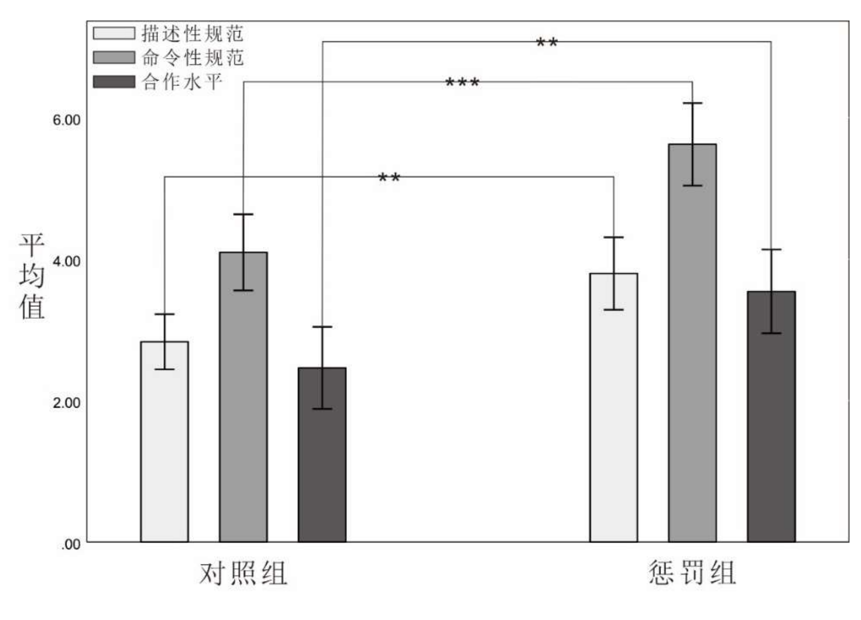
最后，根据黄少安和张苏（2013）对合作所下定义：合作是自己付出成本而使其他人或者公共物品受益的行为，我们用上述 3）和 4）项数字分别表示被试在两种博弈情形下的合作水平（在独裁者博弈中，合作意味着使对方受益；而在公共物品博弈中，合作意味着自己的行为提高了公共物品的产出），数字越大表示合作水平越高。

3.3 结果与讨论

我们首先使用规范激活水平的第一种操作定义进行了统计分析，结果发现，不同性别（$t = 0.07{\sim}1.26, p = 0.209{\sim}0.941$）和专业（$F = 0.18{\sim}1.43, p = 0.236{\sim}0.911$）下描述性规范、命令性规范和合作水平的差异均不显著，年龄与描述性规范、命令性规范与合作水平（$r = 0.03{\sim}0.05, p = 0.540{\sim}0.736$）的相关系数不显著。如图 2 所示，惩罚组被试的描述性规范激活水平（$M = 3.80$，$SD = 2.45$，$n = 90$）显著高于对照组（$M = 2.83$，$SD = 1.85$，$n = 89$）（$t = 2.97$，$p = 0.003$，$d = 0.44$，95%C.I. = [0.15，0.74]）；惩罚组被试的命令性规范激活水平（$M = 5.62$，$SD = 2.79$）显著高于对照组（$M = 4.10$，$SD = 2.56$）（$t = 3.82$，$p < 0.001$，$d = 0.57$，95%C.I. = [0.27，0.87]）；此外，惩罚组被试在独裁者博弈中的合作水平（$M = 3.55$，$SD = 2.83$）也显著高于对照组（$M = 2.46$，$SD = 2.75$）（$t = 2.59$，$p = 0.009$，$d = 0.39$，95%C.I. = [0.09，0.68]）。上述结果为研究问题 2 提供了初步回答，我们发现第三方惩罚不仅显著激活了违规者的两种社会规范，而且提升了违规者在新情境下的合作水平。在第二阶段的独裁者博弈中，不存在可能实施惩罚的第三方，而且对照组和实验组的唯一的差别就在于实验组被试在第一阶段结束时被提醒过其违规行为受到了惩罚，因而对实验 2 结果的合理解释是第三方惩罚的规范提示功能溢出到了新的情境下，在这种情况下即便不存在惩罚机制，但激活了的社会规范依然可以提升违规者的合作水平。

此外，我们使用规范激活水平的第二种操作定义重复了上述检验过程，并得到了相似的结果：惩罚组被试的描述性规范激活水平显著高于对照组（$t = 4.18$，$p < 0.001$）；惩罚组被试的命令性规范激活水平也显著高于对照组（$t = 4.80$，$p < 0.001$）。上述结果意味着我们的研究结论具有较高的稳健性。

为这种在划分标准上的稳定性具有一定的生物学基础（Wallace et al., 2007）。以初始金额（20 代币）30%计算，6 代币为分界点，也就是说高于 6 代币的分配方案可被认为是一种合作行为。



注： ***$p < 0.001$，**$p < 0.01$

图 2 惩罚组和对照组的规范激活与合作水平

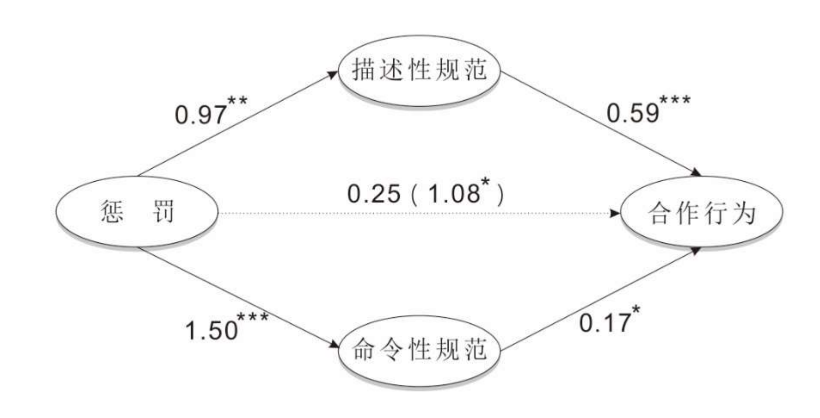
为回答研究问题 4（两种规范在惩罚提升合作中是否具有不同机制？），我们进一步探讨了惩罚影响合作的心理机制，以是否受过惩罚为自变量、描述性规范和命令性规范为中介变量、合作水平为因变量进行中介效应检验。需要说明的是，有研究者指出用偏差校正的非参数百分位 bootstrap 法计算系数乘积的置信区间比 Sobel 法得到的置信区间更精确（方杰, 张敏强, 2012; 温忠麟, 叶宝娟, 2014），因此我们使用 Preacher 和 Hayes（2004）所开发 PROCESS3.5 插件进行中介效应检验（Model 4）。

检验结果如表 2 所示：M_1 和 M_2 中惩罚对两种规范都有显著的影响。与 M_3 相比，M_4 在引入两种规范后 R^2 增加了 0.24，意味着引入两种规范能解释合作行为变异的 24%。进一步分析惩罚通过两种规范对合作行为的间接作用，描述性规范（Effect = 0.57，BootSE = 0.22，BootLLCI = 0.18，BootULCI = 1.06）和命令性规范（Effect = 0.26，BootSE = 0.13，BootLLCI = 0.04，BootULCI = 0.55）的置信区间都不包含 0，这说明两种规范对合作的间接作用都显著；另一方面，是否受过惩罚（Effect = 0.24，$SE = 0.39$，$t = 0.64$，$p = 0.523$，LLCI = -0.51，ULCI = 1.01）置信区间包含 0，这意味着惩罚对合作行为的直接作用不显著。综上所述，惩罚对合作行为的促进作用在很大程度上是通过激活两种社会规范来实现的，两种规范的间接效应占总效应的 77.20%，其中描述性规范的间接效应占 53.08%，命令性规范占 24.12%（图 3），并且两种规范间接效应的大小差异不显著（BootSE = 0.09，BootLLCI = -0.01，BootULCI = 0.30），因此，从实验 2 的结果来看，两种规范在中介惩罚与合作的过程中具有相似的作用机制。

表 2 中介效应的检验

变量	M_1（描述性规范）		M_2（命令性规范）		M_3（合作行为）		M_4（合作行为）	
	系数	*SE*	系数	*SE*	系数	*SE*	系数	*SE*
常数	1.87***	0.51	2.57***	0.63	1.38*	0.66	-0.17	0.62
惩罚	0.97**	0.32	1.53***	0.40	1.08**	0.42	0.25	0.39
描述性规范							0.59***	0.08
命令性规范							0.17*	0.07
模型	R^2	*MSE*	R^2	*MSE*	R^2	*MSE*	R^2	*MSE*
	0.05	4.72	0.08	7.17	0.04	2.79	0.28	5.91

注：括号内为因变量，*** $p < 0.001$，** $p < 0.01$，* $p < 0.05$。



注： *** $p < 0.001$，** $p < 0.01$，* $p < 0.05$。

图 3 描述性与命令性规范的中介作用

同样，我们采用规范激活水平的第二种操作定义进行了稳健性检验，并在上述检验过程中得到了相似的结果：描述性规范间接作用显著（Effect = 0.76，BootSE = 0.21，BootLLCI = 0.38，BootULCI = 1.21）；命令性规范间接作用显著（Effect = 0.33，BootSE = 0.16，BootLLCI = 0.05，BootULCI = 0.68）；直接作用不显著（Effect = -0.01，*SE* = 0.41，*t* = -0.01，*p* = 0.989）。

进一步检验被试在阶段二公共物品博弈中的合作行为可以加深我们对第三方惩罚溢出效应的理解，分析结果显示：惩罚组不仅在与阶段一相同的独裁者博弈中合作水平显著高于对照组，在不同于阶段一的公共物品博弈情境中合作水平（*M* = 5.24，*SD* = 5.70，*n* = 90）同样也显著高于对照组（*M* = 3.76，*SD* = 4.23，*n* = 89）（*t* = 1.97，*p* = 0.050，*d* = 0.30，95%C.I. = [0.001，0.592]）。这说明惩罚的溢出效应不仅体现在与原情境相似的新情境中，也表现在与原情境完全不同的情况下。进一步比较被试在两种博弈情境中合作水平的差异可以让我们

更好地理解溢出效应的机制。由于独裁者博弈和公共物品博弈是两种不同的情境，因此，首先需要将被试的合作水平进行离差标准化，具体而言，根据 Peysakhovich 和 Rand（2016）以及 Rand 等（2014）的建议，我们把独裁者博弈中将 20 代币中的 10 代币分配给对方的方案设为最大值 1（即独裁者博弈中合作水平最高的分配方案），分配给对方 0 代币则为最小值 0（即合作水平最低的分配方案）；类似的，公共物品博弈中将 20 代币全部投入公共账户设为 1（即公共物品博弈中合作水平最高的方案），投入 0 代币（即合作水平最低的方案）则设为 0。分析结果显示：对照组在独裁者博弈（$M = 0.19$，$SD = 0.21$，$n = 89$）和公共物品博弈中（$M = 0.24$，$SD = 0.26$，$n = 89$）的合作行为无显著差异（$t = 1.53$，$p = 0.127$），这在一定程度上说明两种博弈范式本身不会影响被试的合作行为；相反，惩罚组在独裁者博弈下的合作水平（$M = 0.36$，$SD = 0.28$，$n = 90$）显著高于公共物品博弈（$M = 0.26$，$SD = 0.28$，$n = 90$）（$t = 2.35$，$p = 0.020$，$d = 0.35$，95%C.I. = [0.06，0.65]）。

上述结果一方面进一步证实了惩罚的溢出效应，另一方面也意味着惩罚通过激活社会规范所带来的合作提升效果虽然可以跨情境迁移，但不同情境下提升效果比相同情境低。这一结果可以通过 Rand 等（2014）所提出的社会启发法假说（social heuristics hypothesis）得到解释：真实生活中个体间的互动往往是非匿名的和重复博弈的（Dreber et al., 2008; Rand et al., 2016），从长远来看合作是更有利的博弈策略，长此以往，人们内化了这种合作规范并直觉性地将之应用到各种情境中去，但新情境的不同会激发个体的有意识思考，而通过这种思考人们会发现对自身利益而言在新的情境中合作未必是最佳选择（Peysakhovich & Rand, 2016），换言之，理性思考会抑制个体在新情境中的合作行为。就实验 2 的结果而言，当被试从第一阶段的独裁者博弈过渡到第二阶段的公共物品博弈时，个体需要进行一定的思考才能理解两者间的相似与不同，而这种理性思考降低了个体在公共物品博弈中的合作水平；与之相反，第二阶段的独裁者博弈与第一阶段的实验范式无本质差异，被试无需进行思考就能做出直觉反应，因此合作水平更高。

4 实验 3：惩罚促进合作的横向溢出效应

实验 2 验证了第三方惩罚在时间维度上的溢出效应，即第三方惩罚通过激活违规者的社会规范而提高了其在后续新情境下的合作水平，即便在新情境下不存在对违规行为的惩罚机制。实验 3 进一步探讨第三方惩罚的规范激活功能是否能溢出到旁观者或潜在违规者身上，即空间维度上的溢出效应，并比较两种规范的影响机制，从而回答研究问题 3 和 4。

4.1 被试

取中等效应量 f = 0.25，α = 0.05，运用 G*Power3.1 进行的功效分析显示最少需要 158 名被试才能达到 95% (1 – β)的统计检验力，而实际参与实验 3 的被试为不同专业的 160 名本科生，其平均年龄为 21.9 ± 1.93 岁，其中女性占比为 42.50%，被试的专业分布为：理工科占 34.38%、社会科学占 28.13%、人文学科占 26.25%、艺术及其他占 11.25%。在实验开始前被试仔细阅读了有关实验的书面说明并签署了知情同意书。

4.2 设计与程序

实验 3 为 2（旁观前 vs. 旁观后）×2（违规组 vs. 规范组）混合设计。实验开始前，被试被告知他们将观看 1 轮由 3 名成员参与的独裁者博弈，而被试需在博弈完成后尽快计算出各个成员的收益。在了解博弈规则后（分配者拥有 20 代币初始金额，并可将 0~10 之间的任一整数金额分配给接受者，接受者无权干预，但第三方可支付 2 代币来扣减不公平分配者的 6 代币），被试被随机平均分入两种实验条件（80 名违规组被试和 80 名规范组被试），所有被试均被要求估计在即将进行的博弈中：1）将 0、1、2…10 代币分配给接受者的被试的百分比；2）赞成将 0、1、2…10 代币分配给接受者的被试的百分比；3）假设自己为分配者，从 0~10 中选择一个整数代表自己愿意分配给接受者的金额，并且被试被明确告知无论其选择如何都不会受到惩罚。被试完成上述估计后，各自从计算机屏幕上观看 1 轮独裁者博弈：违规组被试看到分配者将 20%的初始金额分给了接受者，并且受到了第三方的惩罚；规范组被试看到的分配方案为 5：5，并且分配者没有受到惩罚。接着，被试计算参与博弈成员的收益，并再一次被要求对在刚完成的博弈中 1）、2）和 3）项数字进行估计。

和实验 2 一样，我们用两种方式计算被试的规范激活水平：第一种方式用 1）和 2）数字各自的加权平均值分别代表描述性规范和命令性规范的激活水平；第二种方式用被试估计有多少比例的分配者将（赞成将）20 代币中的 7、8、9 和 10 代币分配给接受者来代表两种规范的激活水平，我们主要采用第一种操作定义来检验研究问题，而采用第二种操作定义作为稳健性检验。此外，我们用 3）项数字表示被试的合作水平。完成上述步骤后，实验者宣布实验结束，并向被试解释实验设计与目的并支付报酬。实验报酬为 10 元出场费加上随机抽取一种被试的分配方案所产生的代币数。

4.3 结果与讨论

我们首先使用规范激活水平的第一种操作定义进行了统计分析，结果发现，不同性别（t = 0.45~1.51, p = 0.133~0.652）和专业（F = 0.08~1.04, p = 0.374~0.972）下描述性规范、命令性规范和合作水平的差异均不显著，年龄与描述性规范、命令性规范、合作水平（r = -0.05~-

0.03, $p = 0.420$~0.602）的相关系数不显著。以分组（违规组、规范组）和轮次（旁观前、旁观后）做二因素混合设计的方差分析，结果如表 3 所示：分组和轮次的主效应都显著，两者的交互作用也显著。多重比较结果如图 4 所示，违规组被试在旁观惩罚行为后的合作水平（$M = 4.54$，$SD = 2.59$，$n = 80$）显著高于旁观前水平（$M = 2.30$，$SD = 2.37$，$n = 80$）（$SE = 0.42$，$p < 0.001$，95%C.I. = [1.42，3.06]），也显著高于规范组旁观后水平（$M = 2.87$，$SD = 2.73$，$n = 80$）（$SE = 0.42$，$p < 0.001$，95%C.I. = [0.85，2.49]）；规范组旁观前（$M = 2.80$，$SD = 2.82$，$n = 80$）和旁观后无显著差异（$SE = 0.42$，$p = 0.855$，95%C.I. = [-0.90，0.74]）；旁观前两组也无显著差异（$SE = 0.42$，$p = 0.235$，95%C.I. = [-0.32，1.31]）。上述结果表明旁观惩罚行为显著提升了旁观者的合作水平，也就是说，惩罚提升合作的效应的确能溢出到旁观者身上，并且这种溢出效应并非是重复测量引起，因为规范组旁观前后合作水平并无显著变化。这为研究问题 3 提供了肯定的回答。

值得一提的是，两组被试所观察的内容本质上是同一规范的两个面向：遵守规范所以没有遭到惩罚（规范组）或违反规范所以遭受惩罚（违规组），然而这两种不同的呈现方式却产生了截然不同的效果，这在一定程度上暗示，比起展示人们的规范行为来，展示遭受惩罚的违规行为似乎更能让人们意识到社会规范的存在，进而更有效地改变人们的行为模式。这个结果从侧面呼应了 Cialdini 等（1990）的发现：比起完全没有垃圾的场景来，地面有少量垃圾反而更能激活人们的规范意识并提升其环保行为。这可能是因为违规行为一方面从侧面提醒了人们某种规范的存在，另一方面只有极少量违规行为（Cialdini et al., 1990）或违规行为受罚（本研究）则意味着人们对此普遍持不赞许的态度，因此更能促进人们的合作行为。这一发现对制定旨在加强人们合作行为的政策实践具有一定的启发意义。

表 3 二因素方差分析结果

来源	均方	*F*	显著性	偏 η^2
修正模型	75.98	10.95	0.000	0.09
截距	3129.38	451.16	0.000	0.59
轮次	107.07	15.44	0.000	0.05
分组	27.44	3.96	0.048	0.01
轮次×分组	93.42	13.47	0.000	0.04

$R^2 = 0.094$（调整后 $R^2 = 0.086$）

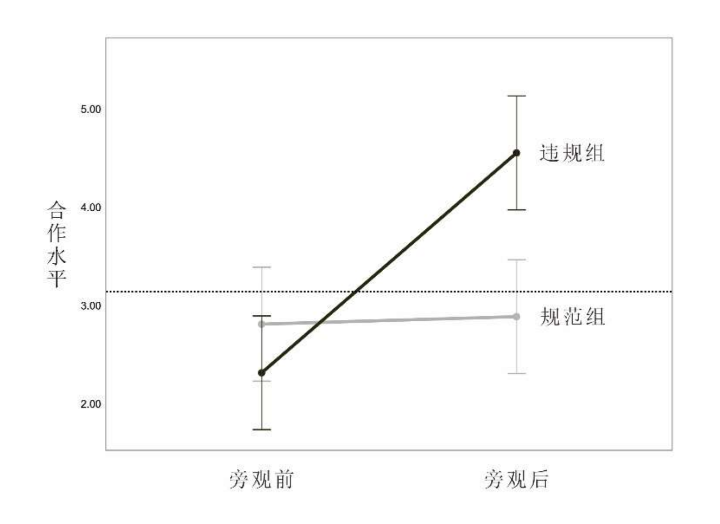
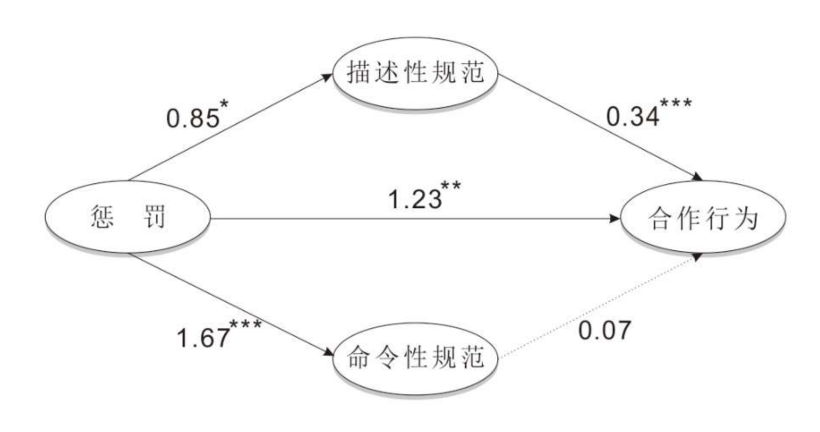


图 4 对合作行为的多重比较

旁观后违规组被试的描述性规范（$M = 3.37$，$SD = 2.20$）显著高于规范组（$M = 2.98$，$SD = 1.89$）（$t = 2.30$，$p = 0.023$，$d = 0.36$，95%C.I. = [0.06，0.73]），违规组被试的命令性规范（$M = 4.97$，$SD = 2.77$）也显著高于规范组（$M = 4.18$，$SD = 2.51$）（$t = 3.32$，$p = 0.001$，$d = 0.52$，95%C.I. = [0.32，0.1.27]），这说明被试合作行为的提高可能是由于旁观惩罚而激活了两种社会规范。进一步以分组（是否看到惩罚）为自变量、描述性规范和命令性规范为中介变量、合作水平为因变量检验社会规范激活是否中介了惩罚与合作行为。bootstrap 检验结果显示规范激活在惩罚与合作行为之间起到部分中介的作用（图 5），其中是否看到惩罚对合作行为的直接作用显著（Effect = 1.23，SE = 0.43，$t = 2.89$，$p = 0.004$，LLCI = 0.39，ULCI = 2.08）；描述性规范（Effect = 0.30，BootSE = 0.16，BootLLCI = 0.04，BootULCI = 0.85）对合作行为的间接作用显著；但命令性规范（Effect = 0.13，BootSE = 0.14，BootLLCI = -0.16，BootULCI = 0.40）对合作的间接作用不显著，并且这种不显著主要体现在“命令性规范→合作”这一路径，也就是说惩罚显著地影响了命令性规范的激活水平，但命令性规范的激活却无法显著改变被试的合作行为。



注：$^{***}p < 0.001$，$^{**}p < 0.01$，$^{*}p < 0.05$。

图 5 描述性规范的中介作用

采用规范激活水平的第二种操作定义进行的稳健性检验同样得到了类似的结果：旁观后违规者的描述性规范（$t = 3.96$，$p < 0.001$）和命令性规范（$t = 4.89$，$p < 0.001$）的激活水平均显著高于规范组，并且描述性规范的间接作用显著（Effect = 0.41，BootSE = 0.19，BootLLCI = 0.10，BootULCI = 0.83）；命令性规范的间接作用不显著（Effect = 0.30，BootSE = 0.20，BootLLCI = -0.06，BootULCI = 0.72），直接作用显著（Effect = 1.43，$SE = 0.49$，$t = 2.90$，$p = 0.004$）。从实验 3 的结果来看，无论采用哪种操作定义，两种规范在中介惩罚与合作中的作用机制似乎存在显著差异，这与实验 2 形成了鲜明对比。

比较描述性规范和命令性规范这两条路径，我们看到实验操作确实同时激活了这两种规范，对两种规范在实验操作前后的平均数差异检验也验证了这一点，两者间的差别主要体现在激活后的描述性规范提升了被试的合作水平，但命令性规范却未能起到类似作用。对上述结果的一种解释是在大部分情况下人们更容易受到描述性规范的影响（陈思静等, 2015; Cialdini et al., 1991），因为描述性规范涉及的是事实判断（人们是怎么做的？），而命令性规范涉及价值判断（人们认为应该怎么做？），个体对事实判断的信息处理速度要高于对价值判断的处理（Deutsch & Gerard, 1955）。进一步比较实验 2 和 3 的结果，可以看到一个明显的差异是在实验 2 中描述性规范和命令性规范的中介效应均显著，且无显著差异，尽管单纯从数字上来看，前者的效应略高于后者，而在实验 3 中描述性规范的中介作用显著，而命令性规范不显著，我们推测这可能是因为两个实验中被试的个人卷入度有所不同：在实验 2 中，被试在第一阶段亲身经历了惩罚，而在实验 3 中被试仅仅旁观了他人受罚，因此可以合理地推测被试在实验 2 中的个人卷入度更高。Petty 和 Cacioppo（1986）指出，当个人卷入度较高时，命令性规范对行为的作用更为明显，这一观点可以解释实验 2 和 3 的差异：由于实验 2 中被试的卷入度更高，因此命令性规范对合作行为的作用也就更为明显，而在实验 3 中低个人卷入度导致命令性规范的影响不显著。

5 总讨论

5.1 研究意义

大量文献探讨了第三方惩罚抑制违规、促进合作的作用（e.g., Fehr & Gächter, 2002; Grimalda et al., 2016; Halevy & Halali, 2015），然而，这种作用是如何产生的这一问题受到的关注相对较少，且现有文献多立足于经济学视角，认为惩罚对违规收益结构的改变是上述作

用的核心机制（韦倩, 姜树广, 2013; Carpenter & Matthews, 2004; Nelissen & Mulder, 2013; Rand et al., 2010）。这一解释恰恰有悖于近年来行为经济学的重要发现：经济人原则在决策过程中并不总是发挥作用（Kahneman, 2011; Thaler, 2016），除非我们先入为主地假定违规者恰好总是理性的经济人。有别于经济学视角，陈思静等（2015）以及 Chen 等（2020）将第三方惩罚视为一种规范提示的手段，换言之，第三方惩罚通过激活了个体内化于心中的合作规范（Rand et al., 2014）来提升其合作水平，而无需涉及个体的经济利益。然而，要在严格意义上得出上述结论，我们就必须排除惩罚对收益的影响，因为在有关第三方惩罚的主流研究中，惩罚总是会降低被试的收益。从上述逻辑出发，实验 1 首次在控制惩罚的经济效应后检验了惩罚的规范提示功能，结果发现，即使惩罚造成的损失小于违规行为带来的收益，第三方惩罚依然能显著抑制违规行为并提升合作水平，换句话说，即便是违规者其行为也未必总是遵循经济人假设。古希腊哲学家苏格拉底的一个著名观点是，人们因为无知而作恶（汪子嵩等, 2004）。本文部分地证实了苏格拉底的智慧：很多时候人们违规只是因为没有意识到某种规范的存在，激活人们的规范意识就能显著降低其自私行为，而第三方惩罚是激活人们规范的重要手段之一。上述发现的一个实践意义是，相比于其他规范激活手段，表现为扣减违规者报酬的经济惩罚可能是低效的，因为经济惩罚需付出成本，而扣除惩罚成本后集体的净收益有可能反而更低了（Dreber et al., 2008）。因此，在政策实践中，我们需识别哪些情境下违规是因为缺少规范意识而哪些是纯粹为了获取个人利益，不区分违规动机而一刀切地实施惩罚可能反而降低了社会的运行效率。

其次，实验 1 的结果还可以解释以往文献的若干发现。Rand 等（2009）以及 Fehr 和 Rockenbach（2003）发现，惩罚动机的合理程度可以极大地影响受罚者的合作行为。纯粹的经济视角并不能完全解释上述现象，但如果我们将第三方惩罚视为规范提示的手段，上述问题便迎刃而解：作为规范提示的惩罚自身必须符合某种规范，也就是说，必须具备某种道德合法性，违反规范的惩罚显然不可能具有规范提示的作用，因而也就失去了促进合作的积极作用。上述观点的一个推论是如果惩罚完全不具备经济功能，那么我们可以在很大程度上排除惩罚的不合理动机（如惩罚是为了提高自身的相对优势），在这种情况下，按照实验 1 的结果，我们应该能观察到这类惩罚对合作同样具有促进作用。事实上，确实有研究者注意到，面对违规行为，他人的言语责备（也有学者将言语责备称为社会惩罚或道德惩罚）就能起到类似的作用（Noussair & Tucker, 2005），而无需对违规者造成具体的金钱或物质损失，甚至比以降低经济收益为目标的惩罚效果更好（Wu et al., 2016）。实验 1 的结果可以解释上述现象：尽管言语责备并未改变惩罚的收益，但和第三方惩罚类似，言语责备起到了提示违规者

存在某种规范的作用，同时言语责备在很大程度上排除了为自身牟利的非法动机，从而有效地降低了违规者的自私行为。当然，也有研究者认为言语责备的作用同样可以从经济角度来解释，比如 van den Berg 等（2012）认为成本表现为多种形式，言语责备尽管未必会提高违规行为的金钱成本，但可能提高了违规者在人际关系方面的成本，因此实际上依然减少了违规者的收益。然而，实验室环境中的言语责备往往程度较轻，如“我认为你的分配方案不公平”（Nelissen & Mulder, 2013）或“某某人只关心自己”等（崔丽莹等, 2017），且经常发生在匿名环境中（陈思静, 徐烨超, 2020），因而似乎很难认为匿名状态下上述言语能对违规者的人际利益造成实质性损害。综上所述，我们认为实验 1 的结果可以更好地解释言语责备对合作的提升作用。

第三，更为重要的是，本文基于社会规范视角提出了第三方惩罚是如何维持人类社会长时间、大规模的合作。惩罚降低收益的经济学观点无法解释上述现象，因为完全理性的个体曾经受罚的经历不足以使其在新情境下表现更好，除非在新情境下依然存在惩罚机制，然而正如 Shreedhar 等（2018）指出，无处不在的惩罚会极大提高社会运行成本。而我们在实验 2 和 3 中所发现的第三方惩罚的两种溢出效应可以很好地解释为什么第三方惩罚可以维持广泛的合作行为：实验 2 表明第三方惩罚的规范激活作用不仅抑制了被试在当前博弈情境中的违规行为，还进一步提高了被试在后续其他博弈情境中的合作水平，我们把上述作用称之为“纵向溢出效应”；而实验 3 的结果显示惩罚的溢出效应不仅发生在不同情境下的同一个体身上，也同样发生在了旁观而非参与博弈的个体身上，即人们只要作为旁观者观察到了违规者受到惩罚，那么惩罚的规范激活作用就能发挥作用，相应的，上述效应或可称为“横向溢出效应”。上述结果意味着，维持大规模合作并不需要无处不在的惩罚，因为某次特定惩罚的效果并不仅仅体现在当下，还可以在时间和空间维度持续发挥作用：惩罚在很大程度上是一种规范提示，因而受罚者或目睹受罚的旁观者通过激活自身内在的规范而抑制了潜在的违规冲动，并使合作水平在一定范围内保持在较高水平，而无需外在的惩罚者时时监督违规行为并处以惩罚。综上所述，本研究的另一意义在于本文所发现的第三方惩罚的两种溢出效应为理解惩罚如何维持人类社会的广泛合作提供了新的理论思路。需要指出的是，我们对上述实验结果的解释并不是唯一的，Gintis 和 Fehr（2012）提出了另一种解释：惩罚对合作的提升仍然依赖于惩罚对违规收益的降低作用，只不过惩罚无需对违规者造成实际的损失，只要违规者担心惩罚有可能给他们造成损失，惩罚就能发挥积极作用。上述观点的确可以从成本-收益的经济学角度来解释为什么少数几次惩罚就可以维持大规模的合作，然而，为了排除这一竞争性假设，我们在实验 2 和 3 中均明确地告知被试无论他/她是否违规，都不会受

到任何惩罚。这一实验设定在相当程度上排除了是被试对潜在惩罚的担心而提高了合作水平。换言之，实验 2 和 3 的结果更加支持惩罚的规范提示解释，而非惩罚的威慑解释。当然，合作作为社会科学中最大的谜题之一（Bear & Rand, 2016），可能并不存在单一的解释，也就是说，我们基于社会规范的解释与 Gintis 和 Fehr（2012）的理论可能并非相互排斥，而是相互补充，从而为合作的演化这一难题提供更完整的答案。

5.2 研究不足

尽管取得了若干有意义的结果，但本研究尚存在一定的不足之处。首先，本研究采用了作为主流研究范式的经济惩罚，即惩罚成本和违规成本均表现为金钱成本，这种设定有利于研究者得出相对清晰的结论，但现实生活中的成本往往表现为多种形式（Guala, 2012），且不同形式的成本会对结果产生不同影响（陈思静等, 2020）。当成本表现为非金钱形式时（如个体的时间、精力或人际资源等），本文的结论是否依然成立是一个值得进一步探索的问题。其次，虽然本研究明确地发现了第三方惩罚的两种溢出效应，但由于前后相隔时间较短（不超过 1 个小时），因此，我们很难确定当受罚或旁观受罚的经历与下一轮博弈间隔较久时（如一周或一个月）这种溢出效应是否依然存在，在较长的时间跨度内展开上述实验有助于提高本研究结论的说服力。再次，在实验 3 中我们比较了规范的两种呈现方式（遵守规范而不受罚 vs. 违反规范而受罚）对被试合作水平的影响，从更为广阔的理论视角来看，更有意义的比较可能是“遵守规范得到奖赏”和“违反规范受到惩罚”这两种展示方式对合作与规范激活的影响，但由于本文的研究焦点在于惩罚对抑制违规、促进合作的作用，因而未能对上述比较做出分析，未来研究可进一步对此进行探索。最后，我们在实验中注意到，惩罚的溢出效应在迁移过程中发生了损耗，这在一定程度上暗示第三方惩罚对维持大规模的合作是有一定界限的，这一结果符合先前研究者的观察，即第三方惩罚的上述作用随着社群规模的扩大而逐渐减弱（Greif, 1993），同时也表明，仅仅依靠自下而上的第三方惩罚似乎还不足以彻底解释人类个体间的广泛合作，而引入其他机制如自上而下的群集惩罚（pool punishment）（Baldassarri & Grossman, 2011）或协调惩罚（coordinated punishment）（韦倩等, 2019）或许能帮助我们更好地理解人类社会的合作现象。

参考文献

Alkan, H. I. (2020). A challenge to homo economicus: Behavioral economics. In I. Akansel (Ed.), *Examining the relationship between economics and philosophy* (pp. 176–195). Hershey, PA: IGI Global.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11023–11027.

Balliet, D., Mulder, L. B., & van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(4), 936–941.

Bicchieri, C., Dimant, E., & Xiao, E. T. (2018). *Deviant or wrong? The effects of norm information on the efficacy of punishment* (PPE Working Papers 0016). Philadelphia, PA: Philosophy, Politics and Economics of University of Pennsylvania.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191–208.

Bingham, P. M. (1999). Human uniqueness: A general theory. *The Quarterly Review of Biology*, *74*(2), 133–169.

Camerer, C. F., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science*, *311*(5757), 47–52.

Carpenter, J. P., & Matthews, P. H. (2004). Why punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, *14*(4), 407–429.

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PloS One*, *15*(3), e0229510.

Chen, S. J., He, Q., & Ma, J. H. (2015). The influence of third-party punishment on cooperation: An explanation of social norm activation. *Acta Psychologica Sinica*, *47*(3), 389–405.

[陈思静, 何铨, 马剑虹. (2015). 第三方惩罚对合作行为的影响: 基于社会规范激活的解释. *心理学报*, *47*(3), 389–405.]

Chen, S. J., Hu, H. M., & Yang, S. S. (2020). Payment vs. retaliation: Impact of cost form on third-party punishment. *Journal of Psychological Science*, *43*(2), 416–422

[陈思静, 胡华敏, 杨莎莎. (2020). 支付与报复: 成本形式对第三方惩罚的影响. *心理科学*, *43*(2), 416–422]

Chen, S. J., & Xu, Y. C. (2020). Warmth and competence: Impact of third-party punishment on punishers' reputation. *Acta Psychologica Sinica*, *52*(12), 1436–1451.

[陈思静, 徐烨超. (2020). "仁者"还是"智者": 第三方惩罚对惩罚者声誉的影响. *心理学报*, *52*(12), 1436–1451.]

Cialdini, B., Kallgren, A., & Reno, R. (1991). A focus theory of normative conduct. *Advances in Experimental Social Psychology*, *24*, 201–234.

Cialdini, B.., Reno, R., & Kallgren, A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026.

Cialdini, B., & Trost, M. (1998). Social influence: Social norms, conformity, and compliance. In T. Gilbert, T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2, 151–192.). Boston, MA: McGraw-Hill.

Csukly, G., Polgár, P., Tombor, L., Réthelyi, J., & Kéri, S. (2011). Are patients with schizophrenia rational maximizers? Evidence from an ultimatum game study. *Psychiatry Research*, *187*(1-2), 11–17.

Cui, L. Y., He, X., Luo, J. L., Huang, X. J., Cao, W. J., & Chen, X. M. (2017). The effects of moral punishment and

relationship punishment on junior middle school students' cooperation behaviors in public goods dilemma. *Acta Psychologica Sinica*, *49*(10), 1322–1333.
[崔丽莹, 何幸, 罗俊龙, 黄晓娇, 曹玮佳, 陈晓梅. (2017). 道德与关系惩罚对初中生公共物品困境中合作行为的影响. *心理学报*, *49*(10), 1322–1333.]
de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, *84*, 103800.
de Kwaadsteniet, E. W., van Dijk, E., Wit, A., de Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin*, *33*(12), 1648–1660.
Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636.
Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351.
Fang, J., & Zhang, M. Q. (2012). Assessing point and interval estimation for the mediating effect: Distribution of the product, nonparametric Bootstrap and Markov Chain Monte Carlo methods. *Acta Psychologica Sinica*, *44*(10), 1408–1420.
[方杰, 张敏强. (2012). 中介效应的点估计和区间估计: 乘积分布法、非参数 Bootstrap 和 MCMC 法. *心理学报*, *44*(10), 1408–1420.]
Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.
Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137–140.
Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468.
Fehr, E., & Williams, T. (2018). *Social norms, endogenous sorting and the culture of cooperation*. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.
Forquesato, P. (2016). Social norms of work ethic and incentives in organizations. *Journal of Economic Behavior & Organization*, *128*, 231–250.
Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510–1510.
Gintis, H., & Fehr, E. (2012). The social structure of cooperation and punishment. *Behavioral and Brain Sciences*, *35*(1), 28–29.
Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*(1), 103–119.
Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *The American Economic Review*, 525–548.
Grimalda, G., Pondorfer, A., & Tracer, D. P. (2016). Social image concerns promote cooperation more than altruistic punishment. *Nature communications*, *7*, 12288.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–15.

Halevy, N., & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(22), 6937–6942.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Huang, S. A., & Zhang, S. (2013). How did cooperative behavior evolve: A summary and review. *Social Sciences in China*, *7*, 79–91.

[黄少安, 张苏. (2013). 人类的合作及其演进: 研究综述和评论. *中国社会科学*, *7*, 79–91.]

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Lois, G., & Wessa, M. (2019). Creating sanctioning norms in the lab: The influence of descriptive norms in third-party punishment. *Social Influence*, *14*(2), 50–63.

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147–151.

Nelissen, R. M., & Mulder, L. B. (2013). What makes a sanction "stick"? The effects of financial and social sanctions on norm compliance. *Social Influence*, *8*(1), 70–80.

Noussair, C., & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, *3*(3), 649–660.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, *194*(4), 561–574.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York, NY: Springer

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology and Evolution*, *30*(2), 98–103.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*(9), 1192–1206.

Rand, D. G., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, *265*(4), 624–632.

Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General*, *145*(4), 389–396.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272–1275.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677.

Shreedhar, G., Tavoni, A., & Marchiori, C. (2018). *Monitoring and punishment networks in a common-pool resource dilemma: Experimental evidence* (GRI Working Papers 292). London, England: Grantham Research Institute on Climate and the Environment.

Sood, S., Kostizak, K., Lapsansky, C., Cronin, C., Stevens, S., Jubero, M., ... & Obregon, R. (2020). ACT: An evidence-based macro framework to examine how communication approaches can change social norms around Female Genital Mutilation. *Frontiers in Communication*, *5*, 29.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review*, *106*(7), 1577–1600.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

van den Berg, P., Molleman, L., & Weissing, F. J. (2012). The social costs of punishment. *Behavioral and Brain Sciences*, *35*(1), 42–43.

Voisin, D., Girandola, F., David, M., & Aim, M. A. (2016). Self-affirmation and an incongruent drinking norm: Alcohol abuse prevention messages targeting young people. *Self and Identity*, *15*(3), 262–282.

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(40), 15631–15634.

Wang, Z. S, Fan, M. S., Chen, C. F., & Yao, J. H. (2004). *A History of Greek Philosophy (Volume 2)*. Beijing: People's Publishing House

[汪子嵩, 范明生, 陈村富, 姚介厚. (2004). *希腊哲学史(第 2 卷)*. 北京: 人民出版社.]

Wei, Q., & Jiang, S. G (2013). How is social cooperative order possible: Exploring mysteries. *Economic Research Journal*, (11), 140–151.

[韦倩, 姜树广. (2013). 社会合作秩序何以可能: 社会科学的基本问题. *经济研究*, (11), 140–151.]

Wei, Q., Sun, R. Q., Jiang, S. G., & Ye, H. (2019). Coordinated punishment and the evolution of human cooperation. *Economic Research Journal*, (7), 174–187.

[韦倩, 孙瑞琪, 姜树广, 叶航. (2019). 协调性惩罚与人类合作的演化. *经济研究*, (7), 174–187.]

Wen, Z. L., & Ye, B. J. (2014). Different methods for testing moderated mediation models: Competitors or backups? *Acta Psychologica Sinica*,*46*(5), 714–726.

[温忠麟, 叶宝娟. (2014) 有调节的中介模型检验方法: 竞争还是替补? *心理学报*, *46*(5), 714–726.]

Wu, J., Balliet, D., & van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports*, *6*, 23919.

Xie, D. J., & Su, Y. J. (2019). The evolutionary and cognitive mechanisms of third-party punishment. *Journal of Psychological Science*, *42*(1), 216–222.

[谢东杰, 苏彦捷. (2019). 第三方惩罚的演化与认知机制. *心理科学*, *42*(1), 216–222.]

Spillover effects of third-party punishment on cooperation: A norm-based explanation

CHEN Sijing[1] , XING Yilin[1] , WENG Yijing[1] , LI Chang[2]

([1]*School of Economics and Management*, *Zhejiang University of Science and Technology*, *Hangzhou*, 310023, *China*) ([2]*School of Business Administration*, *Zhejiang Gongshang University*, *Hangzhou*, 310018, *China*)

Abstract

A large body of experimental evidence demonstrates that in presence of third-party punishers, cooperators can gain higher payoffs than defectors. As a result, third-party punishment (TPP) that changes the payoff structure of defectors is believed to be a key in promoting cooperation. However, this rationale is contrary to an important finding in behavioral economics: individuals are not

necessarily rational decision makers and do not have purely self-regarding preferences. This contradiction raises an interesting question: can this finding also be applied to defectors? We aim to explore this question through three experiments.

In Experiment 1, 240 undergraduates participated in a Public Goods Game and were divided randomly into three conditions: control condition (CC), low defection cost condition (LC), and high defection cost condition (HC). In each round of the game, participants in CC decided whether to contribute 10 tokens from the initial endowment to the public account. All the tokens contributed to the public account were doubled and evenly allocated to all group members. Participants who retained 10 tokens needed to pay a tax of 1 token. The procedures in LC and HC were identical to that in CC. An exception is that in LC and HC, independent punishers could discipline defectors by paying 5 tokens to reduce the payoff of defectors by 1 token in LC and 10 tokens in HC. In Experiment 2, 179 participants who defected in Stage 1 were selected as sample in Stage 2 and were divided randomly into two conditions: CC (89 participants) and punishment condition (PC, 90 participants). Participants in PC were told they had been punished in Stage 1, whereas those in CC received no feedback. All participants' levels of norm activation and cooperation in different games were then measured. Experiment 2 was replicated in Experiment 3, where the participants were not game players but spectators, and their levels of norm activation and cooperation were measured before and after the game. The participants in defection condition observed a defection and the consequent punishment, whereas those in norm condition observed a fair offer and no punishment.

In Experiment 1, the defection cost in LC was lower than that in CC, so participants in LC had a stronger incentive to defect. However, the results revealed a significantly higher cooperation level in LC. A plausible explanation is that the defection cost in form of punishment served as a norm reminder, but cost in form of tax lacked this function, implying that even defectors are not necessarily benefit maximizers. The results of Experiment 2 confirmed this explanation: compared with unpunished defectors, the punished ones manifested a higher level of norm activation. The bootstrap analysis showed that the norm activation completely mediated TPP and cooperation. Experiment 2 also found a spillover effect of TPP: the punished defectors still demonstrated a high cooperation in a new different game where the sanction was absent. Finally, Experiment 3 found another spillover effect of TPP: bystanders who did not experience the punishment in person but witnessed it showed a significantly higher cooperation in subsequent interactions.

In conclusion, oftentimes, people defect simply because they are unaware of the existence of a certain norm, and activating people's norms through TPP can significantly reduce their selfish behaviors. In addition to being an economic means to reduce defectors' payoff, TPP serves as a norm reminder. The two spillover effects found in this study suggest that TPP as a means of norm activation may be more efficient than as an economic means because of its cost-effectiveness. These findings shed new light on the understanding of extensive cooperation among genetically unrelated individuals.

Key words third-party punishment, social norm, cooperation, focus theory, spillover effect